ENGR 350

Project Name: Analysis & Inspection on Cancellation Behavior of Hotel Customers

Phase II

Fırat Tamur

60534

Kutay Eroğlu

59970

DATA:

Data cleaning and transformation were needed. Two columns, ‘country’ and ‘company’ were dropped since they had too many null values and they did not have a high impact in terms of our analysis. Multiple rows which had null values for ‘children’ and ‘country’ were also dropped since those rows were not essential for the analysis. Input and output data format were both csv. Data used in the project was taken from kaggle. After cleaning, processed data consisted of 118898 rows and 30 columns. A database was not used.

PROCESSING:

For visualization, matplotlib and seaborn libraries were used. Specifically, correlation heatmap was used to show affinities between columns. Pie chart was used to contrast total bookings of City Hotel and Resort Hotel. Countplot was used numerous times to find out which values concluded in a higher probability of cancellation. Catplot was utilized to compare probabilities of different hotels’ cancellations. Distplot was used to analyze distribution of several columns. Kdeplot was used to estimate specific column’s correlation with cancelling behavior. Boxplot and several other plot techniques were also used.

PROGRAMMING:

For models, classification algorithms were used. Linear discriminant analysis and neural networks were faster than alternatives. For small portions of our data, different classification algorithms were also used. Scikit learn ad tensorflow modules were used.

CHALLENGE:

Initial data contained cancelled entries which were clustered by themselves. To solve this, at the end of processing we shuffled the rows before shipping them to modeling.

POTENTIAL IMPROVEMENT:

As an improvement, optimized model customers could be created for individual hotels. Even though hotels can already make adjustments to their decision making with the current conclusions of our project, it would be more practical for a hotel to have profitable customer profile. Parameters for models could be fine-tuned to increase prediction accuracy. Also, more data could be collected to acquire better insights.